基于通道权重的顺序精炼 RGB-D 显著检测网络 *

卞华军，王华军，赵赫威

(成都理工大学 网络安全学院，成都 610059)

摘　要：提出了一种新型的用于 RGB-D 显著目标检测的网络框架(SR-Net)。为了有效整合多模态特征的互补性，将深度特征提取作为独立分支，采用卷积块注意模块(CBAM，convolutional block attention module)进行深度特征增强，并整合增强后的深度特征与 RGB 特征的互补信息。为了去除特征冗余，减少背景噪声对预测结果的干扰，在上采样网络中设计了一种顺序精炼网络，即通过整合多层次、多尺度特征的互补性，获取初级全局特征，并采用基于通道权重的初级全局特征权重矩阵获取模块(PFW，primary global feature weight matrix acquisition module)获取初级全局特征的权重矩阵；其次利用获取到的权重矩阵对各层次特征进行精炼，以抑制背景噪声带来的干扰；最后，为了更好的优化整个网络，提出了一种新的损失函数。在四个公共数据集上的实验结果表明，该模型在不同的模型评价指标上均优于近年来 9 种先进方法，获得了优异的性能。

关键词：显著性目标检测；RGB-D；通道权重；顺序精炼

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2021.12.0696

Sequential refined RGB-D saliency detection network based on channel weight

Bian Huajun, Wang Huajun, Zhao Hewei

(*School of network security*, *Chengdu University of Technology, Chengdu* 610059, *China*)

Abstract: This paper proposed a new network framework for RGB-D salient object detection (SR-Net) . In order to effectively integrate the complementarity of multi-model features, this paper took the depth feature extraction as an independent branch, use the Convolutional Block Attention Module(CBAM) to enhance the depth feature, and integrate the complementary information of the enhanced depth feature and RGB feature. Then, in order to remove feature redundancy and reduce the interference of background noise on the prediction results, it proposed a sequential refining network in the up-sampling network, that is, first, the primary global features are obtained by integrating the complementarity of multi-level and multi-scale features, and used the Primary Global Feature Weight Matrix Acquisition Module (PFW) which based on the channel weight to obtains the weight matrix of the primary global feature, and then uses the obtained weight matrix to refine the features of each level to suppress the interference which caused by background noise. Finally, in order to better optimize the whole network, it proposed a new loss function. The experimental results on four public datasets show that the model is superior to nine advanced methods in different model evaluation indexes, and achieves more advanced performance.

Key words: salient object detection; RGB-D; channel weight; sequential refine

0　引言

基于 RGB-D 的显著目标检测(RGB-D SOD)旨在从一对 RGB 图像及深度图像中检测到最具吸引力的部分。在过去的十几年里，显著目标检测(SOD)因可以广泛应用于图像分割[1]，图像编辑[2]以及视频分析[3]等领域的预处理阶段，而备受关注。传统的显著目标检测方法主要依赖于手工制作的低级特征[4, 5,26, 27]来进行显著目标检测，但因缺少对显著目标语义信息的获取而很难在背景比较复杂等情况中取得良好的实验效果。近年来，随着深度学习的快速发展，众多研究工作者开始将卷积神经网络(CNN, convolutional neural networks)应用于 RGB-D SOD 中，并取得良好的实验效果。Li 等[22]首次采用深度神经网络搭建了一个基于多尺度特征的显著性模型；Wu 等[23]提出级联部分解码器模型(cascaded partial decoder, CPD)，将主干网络中较深的特征进行整合，得到初始显著性图，进而通过整体注意力模块细化特征，获得最终的显著性图；Liu 等[24]认为主干网络从浅到深提取多层次特征，生成粗显著图，它定位了显著目标，但失去了轮廓细节，其在 DRCNNNet 中采用 DRCNN 用于从深到浅渲染显著目标。低层侧输出借助于深层侧输出、原始深度线索和粗显著图，可以从多个尺度生成显著对象，从而保留更多地轮廓细节；Wu 等[25]在 MCMF-Net 中提出了一种利用深度数据从相应的几何信息中检测显著目标边界的方法，而不是简单地从深度数据中提取显著目标特征。但是，随着研究工作的不断进行，现仍然存在两种难点亟待解决：一方面是如何有效整合多模态、多尺度及多层次特征的互补性；另一方面如何有效抑制复杂背景噪声带来的干扰，并去除特征中所包含的冗余信息。因此，为了解决以上两种问题，本文提出了一种基于通道权重的顺序精炼 RGB-D 显著目标检测网络(SR-Net)。具体的，在 SR-Net 中，本文采用基于注意力机制的 CBAM (convolutional block attention module)模块增强深度特征并有效整合多模态特征的互补性，并设计一种顺序精炼网络，首先通过多层次、多尺度特征融合以获取初级全局特征(如图 2 所示)，并采用基于通道权重的初级全局特征权重矩阵获取模

收稿日期：2021-12-16；**修回日期：**2022-02-21　**基金项目：**四川省人工智能重点实验室项目(2020RYJ02)；模式识别与智能信息处理四川省高校重点实验室(MSSB-2020-10)

作者简介：卞华军(1996-)，男，江苏盐城人，硕士研究生，主要研究方向为计算机视觉、图像处理(969923258@qq.com)；赵赫威(1997-)，男，河北邢台人，硕士研究生，主要研究方向为机器学习、深度学习；王华军，男，四川成都人，博导，博士(后)，主要研究方向为计算机视觉、人工智能、模式识别.

块 (PFW, primary global feature weight matrix acquisition module)获取初级全局特征的权重矩阵并去除冗余信息，再利用获取到的权重矩阵对各层次特征进行精炼，以抑制背景噪声带来的干扰。

RGB 图像包含了显著目标的颜色、纹理等信息，而深度图像可以获得显著目标的结构及空间布局，两者获取到的特征具有互补性。对比仅将深度图像作为 RGB 图像的补充[6]不同，在下采样网络中，本文采用了两个独立的 Resnet-50 骨干网络分支分别进行深度特征和 RGB 特征提取，提取到的深度特征采用基于注意力机制的 CBAM 模块进行深度特征增强，将增强后的深度特征与 RGB 进行互补性特征整合，有效地整合了多模态特征的互补性。

背景噪声会对最终显著目标预测结果造成严重影响，希望在上采样中，可以去除各层次特征中的冗余信息，并采用初级全局特征对各层次特征进行精炼，以强调和增强各层次中的重要信息。综上，本文在上采样网络中，首先初步整合各层次、各尺度特征的互补性，以同时结合低层次特征中包含的纹理信息和高层次特征中包含的语义信息，获取初级全局特征(见图2 中 F_{prd1})，将获取到的初级全局特征输入到PFW 模块，去除冗余信息，并获取初级全局特征的权重矩阵(如图 2 中 *Weights* 所示)，用以精炼各层次特征，降低背景噪声的干扰。

与以往简单整合多模态特征的互补性不同，首先本文采用了两个 Resnet-50 骨干网络分支进行 RGB 和深度特征提取，并采用 CBAM 模块进行深度特征增强，有效整合了多模态特征的互补性。再者，为了去除各层次特征中包含的冗余信息，降低背景噪声带来的干扰，在上采样网络中设计了顺序精炼网络，并设计了基于通道权重的 PFW 模块，去除初级全局特征中的冗余信息，获取初级全局特征的权重矩阵，用于后续精炼各层次特征。如图 2 所示，提出的模型的显著目标预测结果边缘清晰(如图 1 第一行图像所示)，且结构完整(如图 1 第二、三行图像所示)。综上所述，本文的贡献主要如下：

1) 采用一种基于注意力机制 CBAM 模块进行深度特征增强，与以往工作仅将深度特征作为 RGB 特征的补充不同，采用单独 Resnet-50 单独骨干网络分支进行深度特征提取；

2) 设计了一种顺序精炼网络，首先通过整合多层次、多尺度特征，获取初级全局特征，然后采用初级全局特征的权重矩阵去精炼各层次特征，以去除冗余信息；

3) 设计了一种初级全局特征权重矩阵获取模块(PFW)，其基于注意力机制，对获取到的初级全局特征进行特征冗余去除，获取相应权重矩阵，进而用于精炼各层次特征；

4) 为了更好的优化本文设计的整个网络，提出了一种新的损失函数，经实验证明，在新的损失函数的优化下，本文提出的 SR-Net 在四个公共数据集上均获得优秀的实验效果。

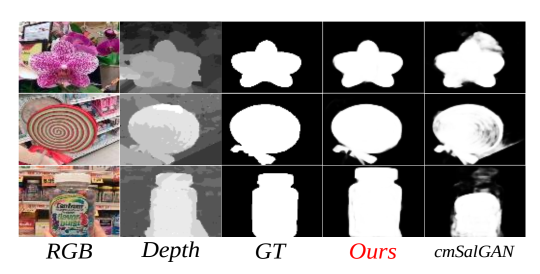


图 1　结果展示

Fig. 1　Result display

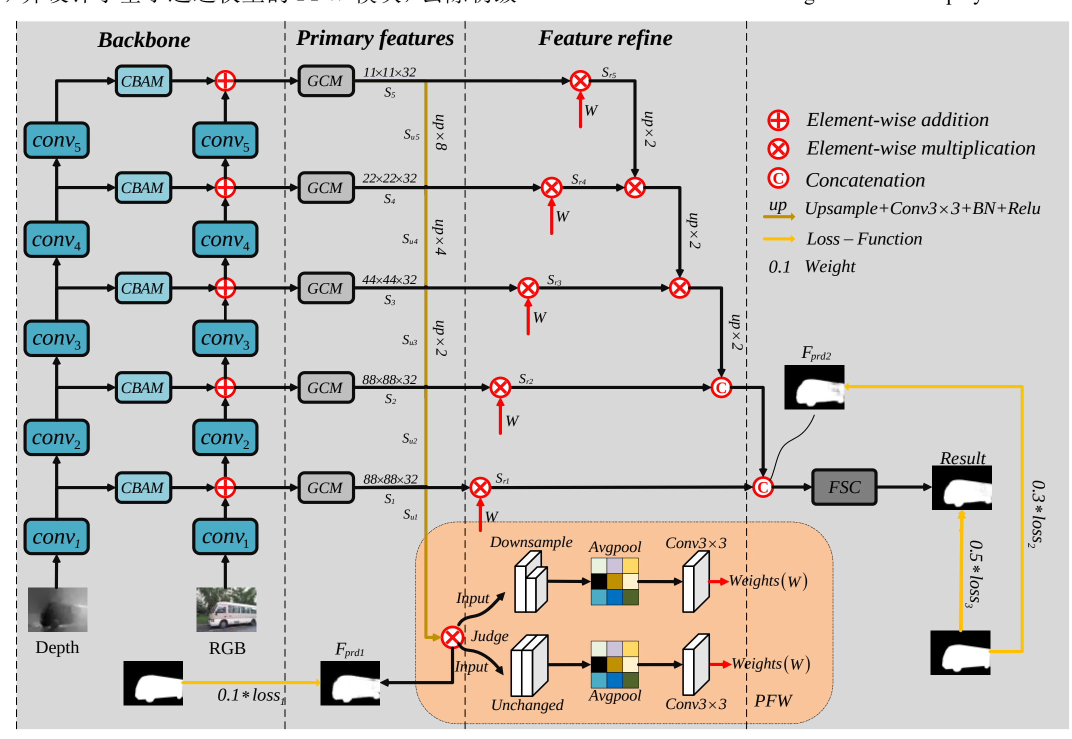


图 2　总体模型架构

Fig. 2　Overall model architecture

1　总体模型架构

如图 2 所示提出了基于通道权重的顺序精炼 RGB-D 显著目标检测网络(SR-Net)。即两个独立的 Resnet-50 特征提取骨干网络分支，一个初级全局特征获取分支，一个基于通道权重采用初级全局特征进行上采样特征精炼分支。具体的，在 Resnet-50 特征提取骨干网络分支中，$conv_i\,(i=1,\ldots,5)$ 分别代表各层特征提取骨干网络，提取到的深度特征会经过深度特征增强模块(CBAM)进行特征增强，随后将增强后的特征与骨干网络提取到的 RGB 特征进行多模态特征融合，获得经过多模态整合后的特征，并输送至上采样网络中。在初级全局特征获取分支中，多模态整合后的特征会首先经过全局上下文获取模块(GCM, global contextual module)及上采样操作来进行上下文信息综合和上采样；其次，整合了经过上述预处理后的多层次及多尺度特征的互补性，获得初级全局特征。在采用初级全局特征进行上采样特征精炼分支中，获取到的

初级全局特征会首先经过基于注意力权重机制的全局特征精炼模块(PFW)，去除初级全局特征的冗余信息，并生成对应权重矩阵(如图 2 中‘ $Weights(W)$ ’所示)，其次，利用生成的权重对各层次特征进行精炼，最后，整合多层次、多尺度精炼后的特征，获取最终的显著目标预测结果。为了更好的优化提出的基于通道权重的顺序精炼网络，本文在网络中的不同层次进行上采样，以获取到的该层次的显著目标预测结果图，并计算子损失函数，特别的，根据该层次对最终显著目标预测结果的影响程度，给予该层次的子损失函数以不同的权重(如图 2 中‘ $0.1*loss_1$ ’所示)。具体的关于整个网络的介绍如下文所述。

1.1　深度特征增强模块(CBAM)

为了有效整合来自 RGB 特征和深度特征的互补性，以往的工作多采用简单的连接方式，例如，级联、对应元素点乘、相加，或仅将深度特征作为 RGB 特征的补充进行多模态特征融合，并未深度考虑由于内在的模态差异及深度特征的冗余性，直接采用简单的方式整合多模态特征融合会带来一些冗余信息和噪声。受研究者工作[7]的启发，本文采用通道注意力机制及空间全局注意力机制构建深度特征增强模块，进而对深度特征进行特征增强。如图 3 所示，将输入的特征图 F_{input} 分别经过 max-pooling 及 avg-pooling，获得关于特征图的各通道权重，然后经过比率变换提取全局通道信息并对应元素相加，获得基于通道注意力机制的特征图 F_{CA}，具体计算过程如下：

$$f_1 = \mathrm{conv}_{c/ratio\to c}(\delta(\mathrm{conv}_{c\to c/ratio}(\mathrm{maxpool}(F_{input})))) \tag{1}$$

$$f_2 = \mathrm{conv}_{c/\,ratio\to c}(\delta(\mathrm{conv}_{c\to c/\,ratio}(\mathrm{avgpool}(F_{input})))) \tag{2}$$

$$F_{CA} = \mathrm{sigmoid}(\mathrm{conv}_{2\to 1}([f_1, f_2])) \tag{3}$$

其中，F_{input} 代表输入特征图，$maxpool$，$avgpool$ 分别代表着全局最大池化和全局平均池化，$conv_{i\to j}$ 代表将通道数由 i 转变到 j 的 1×1 卷积，ratio 代表比例变换，δ 表示 Relu 激活函数，f_1 及 f_2 表示计算过程中的中间过渡变量，F_{CA} 表示经过通道注意力机制精炼后得到的特征图。

随后，将 F_{CA} 分别经过基于空间的 $maxpool$ 及 $avgpool$，获得空间层面上的关于显著目标的权重，然后采用级联进行连接，并通过 7×7 卷积将通道数转换为 1，获得基于空间注意力机制的特征图 F_{SA}，具体的计算过程如下：

$$F_{SA} = \mathrm{sigmoid}(\mathrm{conv}_{2\to 1}[\mathrm{maxpool}(F_{CA}), \mathrm{avgpool}(F_{CA})]) \tag{4}$$

其中，F_{CA} 表示经过通道注意力机制精炼后得到的特征图，$maxpool$ 及 $avgpool$ 分别表示基于空间的全局最大池化和全局平均池化，F_{SA} 表示经过全局注意力精炼后得到的特征图。

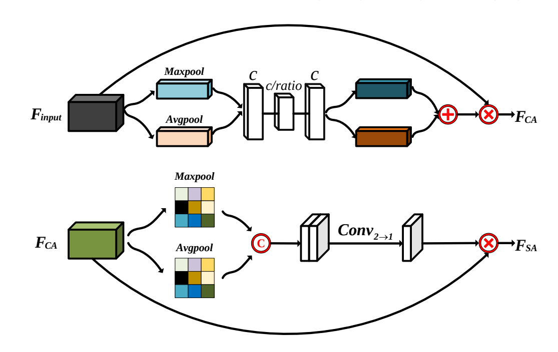


图 3　CBAM 特征增强模块

Fig. 3　CBAM feature enhancement module

1.2　初级特征获取

如图 1 所示，经过深度增强后的特征会和骨干网络提取到的 RGB 特征进行对应元素相加，以整合上下文信息并输送到全局上下文信息获取模块(GCM)，进行上下文信息综合，获得特征 S_i。随后，因多层次、多尺度的特征所包含的关于显著目标的信息具有互补性，有效整合多层次、多尺度特征的互补性，进而获取到的初级全局特征，会包含更多的关于显著目标的主要信息，当使用其进行基于注意力机制的全局信息权重获取时，权重的置信度会更高。基于以上思想，并因各层次特征的尺度不同，首先将各层次的特征经过上采样到相同的尺寸大小(88*88*32)，具体的上采样(up*n)计算过下：

$$S_{ui} = \mathrm{Relu}(BN(\mathrm{conv}_1(upsample*n(S_i)))) \tag{5}$$

其中，S_i 代表通过全局上下文信息获取模块(GCM)去除冗余信息后的特征，$upsample*n$ 代表对 S_i 进行 n 倍的上采样操作，$conv_1$ 代表 3*3 的卷积，BN 表示正则化，Relu 代表 Relu 激活函数，S_{ui} 表示经过上采样后的输出特征。最后，上采样后的各层次特征会进行对应元素相乘，获取初级全局特征，具体的获取初级全局特征的计算过程如下：

$$F_{prd1} = S_{u1}\otimes S_{u2}\otimes S_{u3}\otimes S_{u4}\otimes S_{u5} \tag{6}$$

其中，S_{ui} 表示经过上采样后的输出特征，⊗ 代表对应元素点乘，F_{prd1} 代表获取到的初级全局特征。

1.3　初级全局特征权重矩阵获取模块(PFW)

如图 4 所示，在初级全局特征获取分支中，有效整合了多层次、多尺度特征的互补性，获得初级全局特征 F_{prd1}。因全局特征会包含更多的关于显著目标的重要特征，因此，当采用全局特征指导精炼各层次的特征时，可以去除该层次特征中所包含的冗余信息，并自动选择和增强该特征中所包含的重要特征，降低背景噪声干扰。基于以上思路，提出了初级全局特征权重获取模块(PFW)，具体内容如下所述：

首先，经过初级全局特征获取分支获取到的初级全局特征 F_{prd1} 会根据其即将进行精炼的网络层次进行是否进行下采样判断，值得注意的是，考虑到上采样的过程相较于下采样会引入更多的噪声，在统一不同尺寸的特征时，本文选择将 F_{prd1} 进行下采样，而非对较小尺寸的特征进行上采样。具体的下采样判断的计算公式为：

$$F_{prd1} = \begin{cases} F.\mathrm{interpolate}\,(F_{prd1}), & \text{if size}_{S_i} \mathrel{!}= \text{size}_{F_{prd1}} \\ F_{prd1}, & \text{otherwise} \end{cases} \tag{7}$$

其中，$size_{S_i}$，$size_{F_{prd1}}$ 分别代表各层次特征和初级全局特征的尺寸，F.interpolate 代表基于双线性插值的下采样操作，F_{prd1} 代表经过下采样判断过程后的输出结果。然后，经过下采样后的输出结果 F_{prd1} 均会经过空间层次的全局平均池化，特别的，在这一部分，本文对 F_{prd1} 进行了空间全局平均池化，而非空间全局最大池化，主要原因在于，本文认为最大池化会伴有特殊性及不稳定性，单个通道的权重会对最终整体权重分布造成极大的影响，因此采用空间全局平均池化，可以更加确保整个网络的鲁棒性和准确性。

最后，经过全局平局池化的特征会先后经过 3×3 的卷积和 sigmoid 激活函数，生成最终的关于初级全局特征的权重矩阵，用于后续指导精炼各层次特征。具体的计算过程如下：

$$\mathrm{Weights}(W) = sigmoid(conv_2(savgppol(F_{prd1}))) \tag{8}$$

其中 F_{prd1} 代表经过下采样判断过程后的输出结果，$savgpool$ 代表基于空间的全局平均池化，$sigmoid$ 代表 sigmoid 激活函数，$Weights(W)$ 代表关于初级全局特征的空间权重矩阵。

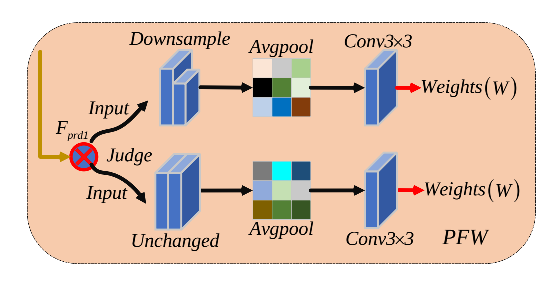


图 4 初级全局特征权重矩阵获取模块

Fig. 4　Primary global feature weight matrix acquisition module

1.4 特征精炼网络

如图 2 所示，因初级全局特征会包含更多的关于显著目标的信息，当用其指导精炼各层次网络的特征，可以去除该层次特征中所包含的冗余信息，并自动选择和增强关键信息。因此，将获取到的初级全局特征的空间权重与各层次特征进行点乘，以获得经过初级全局特征精炼后的各层次特征。随后，按顺序自定向下地整合各层次精炼后的特征，以有效地结合多层次、多尺度特征的互补性，并获得最终的显著目标预测结果。具体的特征精炼过程如下：

$$S_{ri}=W_i\otimes S_i \tag{9}$$

其中，S_i 代表通过全局上下文信息获取模块(GCM)去除冗余信息后的特征，$\otimes$ 代表对应元素点乘，W_i 代表初级全局特征的空间权重，S_i 代表经过初级全局特征指导精炼后的各层次的输出结果。

再者，因经过初级全局特征精炼后的各层次特征所包含的信息不同，为了整合各层次、各尺度特征的互补性，本文自上而下地将各层次特征进行对应元素点乘或级联，为了更清晰地叙述整个整合流程，在这里将输入实例化为 S_{r4} 及 S_{r5}，具体的计算过程为

$$S_{r45}=BN(conv_3(S_{r4}\otimes S_{r5})) \tag{10}$$

其中，S_{r4} 及 S_{r5} 为经过初级全局特征精炼后的各层次特征，S_{r45} 为整合了上述两层特征的互补性后获取到的特征。

最后，将融合了多层次、多尺度后的特征(F_{prd2})上采样到与真值图(GT，Ground Truth)相同尺寸(352×352)，并考虑到直接进行上采样会损失一些细节，并带来噪声，为了解决这一问题，本文采用了一种简单且有效的特征尺寸转换模块(FCS, feature size conversion module)。具体的，FSC 首先采用 1×1 的卷积将特征通道数进行改变，然后，采用残差网络对输入特征图进行上采样，提高信息流通，并防止因网络深度造成的梯度消失和退化问题，具体的计算过程如下：

$$f_3=\mathrm{Re}lu(BN(conv_{96\to 1}(F_{prd2}))) \tag{11}$$

$$\mathrm{Result}=\mathrm{Re}lu(BN(conv_4(f_3))+BN(conv_5(f_3))) \tag{12}$$

其中，F_{prd2} 为上采样网络的最终输出，Relu 代表 Relu 激活函数，f_3 表示中间过渡变量，$conv_4$ 及 $conv_5$ 代表残差网络中采用不同尺寸的卷积层对特征图进行上采样的操作，$Result$ 为整个模型的最终预测结果。

1.5 损失函数

为了更好地训练整个网络，在本文中提出了一种新的损失函数，实验表明，在新的损失函数的优化下，整个模型可以收敛到最低点，最终的显著目标预测结果结构更加完整，边缘更加清晰，损失函数的具体构成如下所述。

如图 2 所示，将初级全局特征、特征精炼分支的输出及最终的显著目标预测结果上采样到与真值图相同尺寸的大小，具体的上采样过程已在 1.2 节式(5)进行了详细介绍，然后对经过上采样后的特征分别进行损失函数计算，损失函数搭建在二元交叉熵损失函数上，二元交叉熵损失函数的计算公式为

$$\ell=G\log S+(1-G)\log(1-S) \tag{13}$$

其中，G 代表真值图，S 代表预测结果图，当计算结果越小时，代表最终的预测结果越贴近真值图。为了更好地让损失函数贴近整个模型的实际运行状态，给予不同层次融合节点的损失函数以不同的权重，以强调随着融合进程，各网络层次的预测结果对最终的显著目标预测结果影响程度，具体的损失函数公式如下所示。

$$\ell_{loss}=0.1\ell_{loss1}+0.3\ell_{loss2}+0.5\ell_{loss3} \tag{14}$$

其中，$\ell_{lossi}(i=1,2,3)$ 分别代表上采样网络的不同融合节点所计算得到的损失函数，ℓ_{loss3} 为对整个模型最后预测输出所计算得到的损失函数，ℓ_{loss} 为总体损失函数。

2 实验及结果分析

在这部分首先对本文中所采用的 4 种公共数据集、5 种评价指标及相应实验细节进行大致介绍，然后会将提出的方法与近年来 9 种先进的模型进行比较，最后，通过一系列的消融实验来证明本文中所提出的一些方法和模块的有效性。

2.1 数据集

为了有效验证 SR-Net 模型的有效性，在 4 个公共数据集上进行了综合实验。即 SIP[8],NJUD[9],NLPR[10],LFSD[11]。其中，SIP[8]包含了通过华为 Meta10 获取到的 929 张高分辨率人物图像，且数据集多集中于现实世界的人物中，NJUD[9]数据集包含了从互联网及 3D 中电影收集到的 1985 张图像，NLPR[10]包含了 1000 张 RGB-D 图像，具有像素级真值图，深度图像是通过 Kinect 在不同照明条件和采集场景下捕获，数据集的图像中可能存在多个显著对象，LFSD[11]包含了 100 张由 Lytro light field camera 相机分别从室内外采集到的分辨率为 360×360 的图像。

2.2 评价指标

为了从定量的角度去评判本文提出的整个模型的好坏，在实验中引入了精准-召回率曲线(PR 曲线)及 5 种评价指标，分别为 F_{max}，F_{ada}，F_β^ω，S_m，MAE。PR 曲线可以通过由一系列精确召回对生成，所获的曲线越接近于(1，1)，越代表模型的预测结果精度越高，具体的精准率(P)和召回率(R)的计算公式为

$$P=\frac{|S'\cap G|}{|S'|},R=\frac{|S'\cap G|}{|G|} \tag{15}$$

其中，G 表示真值图，S' 是根据阈值的预测结果图 S 的二值化掩码。因精准率和召回率有时可能会相互矛盾，因此需要综合考虑，最常用的方法是 F_{max}，即 F_{max} 是精准率和召回率的加权调和平均值，定义为

$$F_{max}=\frac{(1+\beta^2)P\times R}{\beta^2\times P+R} \tag{16}$$

其中，P、R 分别代表精准率和召回率，β^2 代表权重 遵从[12]的建议，本文将 β^2 设置为 0.3 以强调精度。 MAE 表示模型预测结果与真值图的平均像素级误差，当数值越小时，表示模型的预测精度越高。具体计算公式为

$$MAE=\frac{1}{H\times W}\sum_{y=1}^{H}\sum_{x=1}^{W}|S(x,y)-G(x,y)| \tag{17}$$

其中，S 代表模型的预测结果，G 代表真值图，H 及 W 分别代表预测结果图的高度和宽度。

2.3 实验细节

遵从[12],[13]的意见，从 NJUD 及 NLPR 数据集中分别选择 1485 及 700 张图片作为训练集，其剩余图片将与 SIP 及 LFSD 数据集共同作为测试集进行模型测试。本文使用 Resnet-50 作为骨干网络，并使用 Adam 算法进行整个网络的优化，将整个网络在一块 batchsize 设置为 8 的 NVIDIA GeForce RTX 2080Ti GPU 上进行训练，网络初始学习率设置为 1e-4，并使其每隔 60epoch 降低至原来的 0.1 倍，整个网络在 200epoch 停止训练，并保存最好的模型进行测试，整个模型的实验搭建在 Pytorch 平台上。

2.4 与先进的模型比较

在这部分，本文将从定性与定量两种角度将本文提出的 SRNet 与近年来最先进的 9 种模型[14-21]进行比较。为了公平起见，使用作者所给出的源代码进行实验结果复现(如[18])，或直接使用作者给出的该模型的显著目标预测结果。

2.4.1 定性分析

1) 如图 5 所示，本文从 9 种对比模型中随机选取 6 种先进模型同 SR-Net 进行了定性分析。具体的：如图 5 的第一行

图像所示。首先，在对人手及所持物体检测时，众多检测方法，如 CoNet[16]，BiANet[18]，CMWNet[19]，D3Net[20]未能获取到准确的显著目标，并且检测结果中含有大量的噪声。再者，cmSalGAN[14]虽然检测到显著目标的大致轮廓，但缺少了很多边缘细节。相反的，本文的模型能够准确地将人手及所持物体检测出来，并且显著目标的边缘更加清晰，第二行图像同样证明了这一点。

2) 本文提出的 SRNet 能够在多目标情景中，精准检测到显著目标。参见图 5 第三行图像，由于图像中包含了多目标，受多目标的干扰，一些检测方法，如 BBS-Net[15]，CoNet[16]，BiANet[18]，CMWNet[19]，D3Net[20]未能准确检测到主要显著目标，且检测结果中或多或少的包含了噪声。相反的，本文所提出的模型能够精准获取多目标中的显著目标，并且有效减少了噪声干扰，图 5 中第四行图像亦是如此。

3) 本文提出的模型能够在复杂背景下，获取到显著目标。参见图 5 第 6 行图像，由于汽车后部复杂的背景的干扰，一些检测模型未能将整个汽车的完整轮廓进行检测出来，如 CoNet[16]，BiANet[18]。再者，虽然 CMWNet[19]，D3Net[20]获取到了汽车的大致轮廓，但附带了众多的噪声，使得整个检测结果看上去较为杂乱。相反的，如图所示，提出的模型能够完整将汽车检测出来，并且有效地减少了背景噪声带来的干扰，这充分证明了提出的模型同样可以有效应对复杂背景问题。

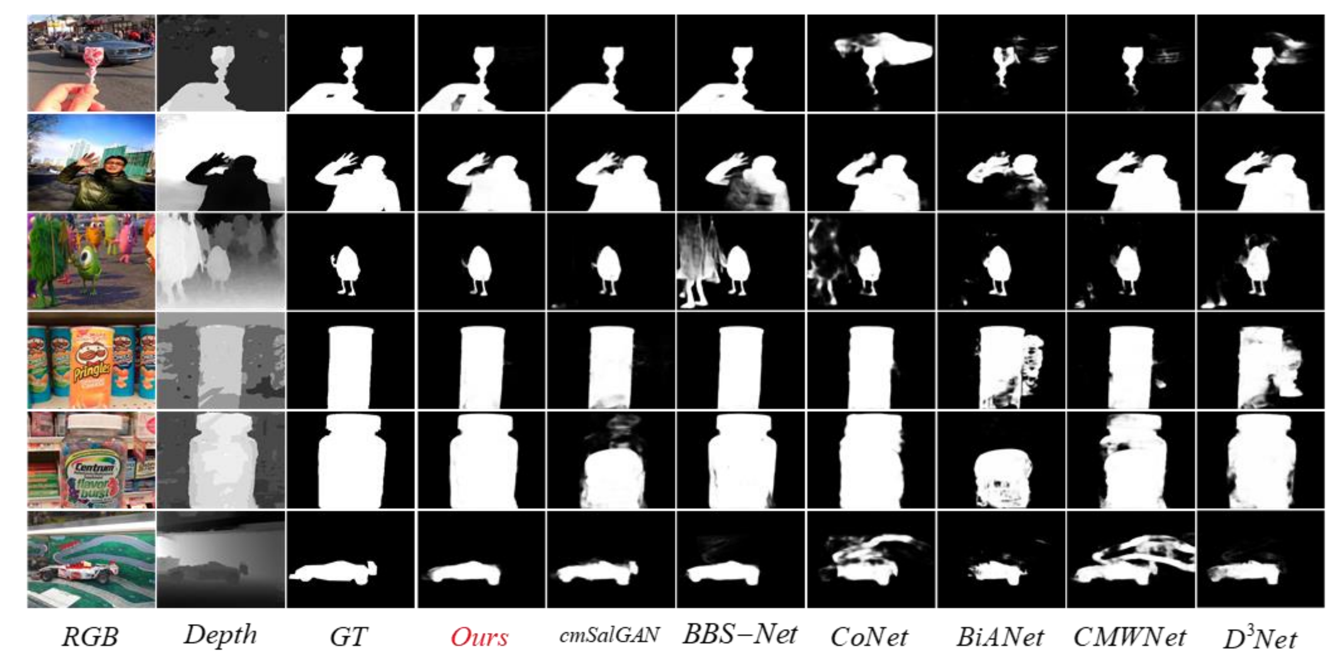


图 5　与其他先进模型视觉效果对比

Fig. 5　Visual effect comparison with other advanced models

2.4.2 定量分析

为了更加直观的展现本文提出的模型的有效性，如表 1 及图 6 所示，从定量角度将模型与 9 种最先进方法在 5 种评价指标及 PR 曲线上进行比较，具体的：

如图 6 所示，本文提出的模型在三个公共数据集(SIP，NJUD,NLPR)上均取得了最高的精准-召回率，仅在 LFSD 数据集上取得次优的结果。再者，如表 1 所示，本文在 5 个评价指标上将模型与对比方法进行定量评估，可以直观得到，在 SIP 及 NLPR 数据集上，本文模型在 5 种评价指标上均优于近年来最先进的方法，与时间维度最近的 cmSalGAN (TMM21)[14]相比，本文提出的模型在四个数据集上均大幅度领先，例如，在 SIP 数据集上，SRNet 相较于 cmSalGAN 在 F_{ada} 及 F_β^ω 指标上分别提高了 2.6%和 3.9%，在 MAE 指标上降低了 17%，这充分证明了提出的模型相较于最新的 cmSalGAN[14]模型,实验效果更加出色。最后，与 9 种对比方法中的相对最优方法，BBS-Net，相比，本文提出的 SRNet 仍然可以取得杰出的实验效果，具体的，SRNet 在 SIP 及 NLPR 数据集上的 5 种评价指标均优于 BBS-Net，仅在 NJUD 及 LFSD 数据集上的一些评价(如 MAE)指标略低于 BBS-Net，这充分证明，本文提出的模型与相对最优方法相比，仍然具有明显优势。

表 1　与其他先进模型定性结果比较

Tab. 1　Comparison of qualitative results with other advanced models

Methods	SIP					NLPR					NJUD					LFSD				
	F_{MAX}	F_{ada}	F_β^ω	E_m	MAE	F_{MAX}	F_{ada}	F_β^ω	E_m	MAE	F_{MAX}	F_{ada}	F_β^ω	E_m	MAE	F_{MAX}	F_{ada}	F_β^ω	E_m	MAE
SRNet(Ours)	**0.905**	**0.875**	**0.834**	**0.918**	**0.053**	**0.927**	**0.885**	**0.881**	**0.957**	**0.023**	**0.932**	**0.901**	**0.885**	**0.923**	**0.035**	**0.882**	**0.859**	**0.812**	**0.8090**	**0.074**
CmSalGAN	0.890	0.849	0.795	0.902	0.064	0.923	0.863	0.855	0.947	0.027	0.910	0.874	0.846	0.907	0.046	0.851	0.831	0.761	0.870	0.097
BBS-Net	0.902	0.872	0.830	0.916	0.055	0.927	0.882	0.879	0.952	0.023	0.931	0.902	0.884	0.924	0.035	0.879	0.858	0.814	0.889	0.072
CoNet	0.883	0.842	0.803	0.909	0.063	0.898	0.848	0.842	0.934	0.031	0.902	0.872	0.849	0.912	0.046	0.877	0.848	0.815	0.896	0.071
$DANet_{vgg16}$	0.901	0.864	0.829	0.916	0.054	0.913	0.871	0.858	0.949	0.028	0.905	0.877	0.853	0.916	0.046	0.871	0.827	0.789	0.827	0.082
$DANet_{vgg19}$	0.892	0.855	0.822	0.914	0.054	0.921	0.875	0.868	0.952	0.027	0.910	0.871	0.857	0.908	0.045	0.871	0.831	0.795	0.874	0.079
BiANet	0.835	0.800	0.739	0.873	0.083	0.893	0.861	0.830	0.940	0.032	0.884	0.849	0.820	0.906	0.055	0.775	0.740	0.675	0.803	0.123
CMWNet	0.890	0.851	0.811	0.907	0.062	0.913	0.859	0.856	0.940	0.029	0.913	0.880	0.857	0.911	0.046	0.900	0.871	0.834	0.891	0.066
D^3Net	0.881	0.835	0.799	0.902	0.063	0.907	0.862	0.849	0.944	0.030	0.909	0.865	0.854	0.913	0.047	0.840	0.801	0.760	0.853	0.095
CPFP)	0.870	0.819	0.788	0.899	0.064	0.888	0.823	0.813	0.924	0.036	0.890	0.837	0.828	0.896	0.053	0.850	0.813	0.772	0.867	0.088

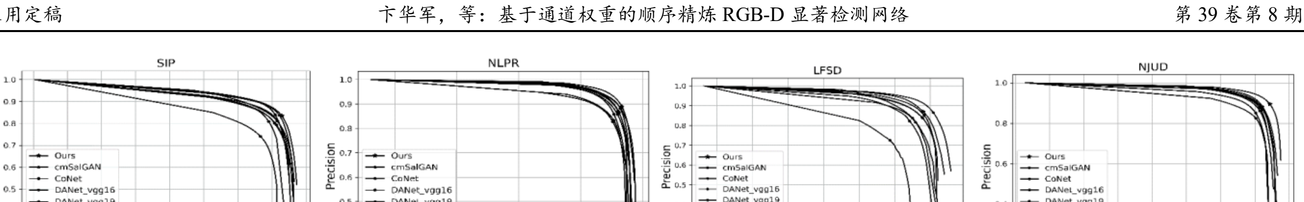


图 6　PR 曲线

Fig. 6　PR curve

2.5　消融实验

在这一部分，将进行消融实验以验证在 SR-Net 中设计的顺序精炼网络、PFW 模块以及损失函数。具体的：

1)为了验证本文提出的顺序精炼网络的有效性，本文将图 2 中的三个融合节点(F_{prd1} , F_{prd2} , Result)分别进行可视化，可视化结果如图 7 所示，可以直观得到，随着顺序精炼网络的进行，在初级全局特征的指导下，图像中的显著目标逐渐完整。并过滤了大部分背景噪声。再者，为了更加充分的证明本文提出的顺序精炼网络的有效性，同样将三个融合节点的输出分别进行定量分析，如表 2 所示，在三个数据集上对三个融合节点进行 5 种模型评价指标测量，实验结果如表 2 所示，可以清晰获得，随着顺序精炼网络的进行，融合节点所获得的显著目标检测结果质量在不断提高。因此，通过视觉与定量两种角度，都完美的验证了本文提出的顺序精炼网络的有效性。

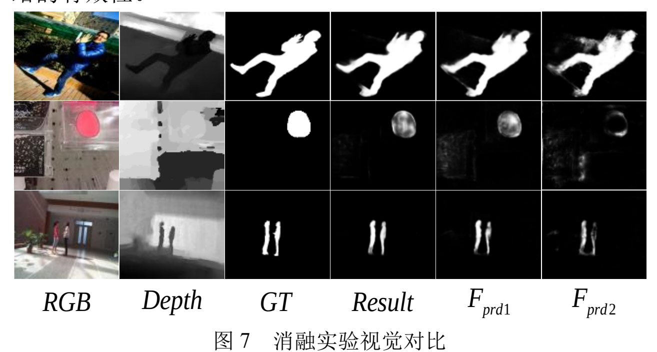


图 7　消融实验视觉对比

Fig. 7　Visual contrast of Ablation Experiment

2)如 1.3 所述，首先获取到初级全局特征，因初级全局特征会包含大量的关于显著目标的主要特征，因此当使用其作为指导特征时，可以精炼和加强被指导特征中所包含的重要特征，并去除冗余信息，因此本文提出 PFW 模块，以去除冗余信息，并获取初级全局特征的权重矩阵，以便于指导融合。为了证明本文提出的 PFW 模块的有效性，将图 2 中的 PFW 去除(对比模型标注为 $SRNet_1$)，初级全局特征仅通过将各层次特征进行对应元素相乘而获得，后续并未通过 PFW 模块去除初级全局特征包含的冗余信息，获取权重矩阵，具体的消融实验结果如表 3 所示。从表中可以获得，在未采用的 PFW 模块的对比模型中，其在三个数据集上的实验结果均低于 SR-Net，且平均降低了 1~2 个百分点，这充分证明了本文在 SR-Net 中所提出的用来获取初级全局特征权重的 PFW 模块的有效性。

3)如 1.5 所述，为了更好地训练整个网络，设计了一种新的损失函数，并给予不同融合节点以不同的权重，进而强调不同融合节点对最终损失函数的影响程度不同。为了验证本文所提出的损失函数的有效性，将所设计的损失函数进行改变，即本文仅计算最终显著目标预测结果的损失函数，并给予权重为 1，而并未计算过程中的融合节点的损失函数，具体的计算公式可以表示为 $\varsigma_{loss} = \varsigma_{loss3}$ 。消融实验结果(对比模型标注为 SRNet2)如表 3 所示。可以获得，在本文设计的损失函数的优化下，本文的实验结果相较于 SRNet2，在三个数据集上均处于全指标领先，领先程度也均处于 1~2 个百分点，这充分证明了，在本文设计的新的损失函数的优化下，可以获得更加精准的显著目标预测结果。

表 2　消融结果

Tab. 2　Ablation result 1

Layer	SIP					NLPR					NJUD				
	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE
Result	0.905	0.875	0.834	0.918	0.053	0.927	0.885	0.881	0.957	0.023	0.932	0.901	0.885	0.923	0.035
$Predict_2$	0.898	0.870	0.817	0.914	0.059	0.920	0.879	0.863	0.954	0.026	0.927	0.895	0.870	0.921	0.039
$Predict_1$	0.893	0.856	0.801	0.912	0.062	0.915	0.858	0.848	0.945	0.029	0.922	0.884	0.860	0.914	0.042

表 3　消融结果 2

Tab. 3　Ablation result 2

Category	SIP					NLPR					NJUD				
	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE	F_{MAX}	F_{ada}	F_{β}^{ω}	E_m	MAE
SRNet	0.905	0.875	0.834	0.918	0.053	0.927	0.885	0.881	0.957	0.023	0.932	0.901	0.885	0.923	0.035
$SRNet_1$	0.896	0.860	0.819	0.914	0.057	0.919	0.869	0.868	0.947	0.026	0.928	0.894	0.880	0.914	0.038
$SRNet_2$	0.903	0.873	0.822	0.911	0.059	0.912	0.868	0.863	0.945	0.027	0.928	0.898	0.879	0.919	0.038

2.6　失败案例

为了促进未来研究工作者对这一领域的研究，在这一部分，将对实验过程中的一些失败案例进行介绍，并给出对该失败案例的一些思路，如图 8 所示，具体的：

1)深度图误导。如图 8 第一行图像中，因深度图像主要突出了第一个玩具，而并未强调后续玩具，促使本文模型及 CoNet[10]在显著目标预测时，只将第一个玩具作为预测结果检测出来，并未识别到后续玩具。第二行图像同样证明了本文的这一观点。

2)与显著目标颜色对比度相近的背景的干扰。如图 8 第三行图像，由于 RGB 图像中的雕塑与背景玩具的颜色十分相近，即使深度图只强调了雕塑，但因 RGB 图像中颜色对比度相近的背景的干扰，Ours,cmSalGAN[14],CoNet[16]在检测过程中，都包含了来自背景的噪声。

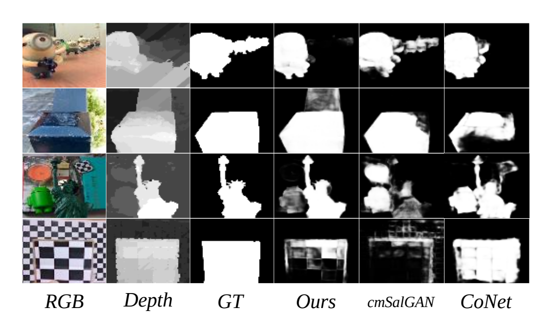


图 8 失败案例

Fig. 8　Failure cases

3　结束语

本文提出一种新型的用于 RGB-D 显著目标检测的网络框架(SR-Net)。为有效整合多模态特征的互补性，将深度特征提取作为独立分支，并采用深度特征模块 CBAM 进行深度特征增强，整合增强后的深度特征与 RGB 特征的互补信息。其次为了去除特征冗余，减少背景噪声对预测结果的干扰，在上采样网络中设计了一种顺序精炼网络，即通过整合多层次、多尺度特征的互补性，获取初级全局特征；采用通过 PFW 模块获取到的初级全局特征的权重矩阵进行各层次特征的精炼；最后提出一种新的损失函数，在四个公共数据集上的实验结果表明该模型在不同的模型评价指标上均优于近年来 9 种先进方法。

参考文献：

[1] Wang Wenguan, Shen Jianbing, Yang Ruigang, *et al.* , Saliency-aware video object segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, pp. 20-33.

[2] Cheng Mingming, Mitra N J, Huang Xing, et al Repfinder: Finding approximately repeated scene elements for image editing, [C]// ACM Transactions on Graphics, 2010: 83: 1-83: 8.

[3] Fan Dengping. , Wang Ww, Cheng Mingming, et al Shifting more attention to video salient object detection [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2019: 8554-8564.

[4] Borji A. and Itti L, State-of-the-art in visual attention modeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012: 185-207.

[5] Borji A. , Saliency prediction in the deep learning era: Successes and limitations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 679-700.

[6] Guo Jingfang, Ren Tongwei, Bei Jia, et al Salient object detection in RGB-D image based on saliency fusion and propagation [C]// Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS) , 2015: 1-5.

[7] Woo S, Park J, and Lee J Y, Cbam: Convolutional block attention module [C]// Proceedings of the European conference on computer vision (ECCV) , 2018: 3-19.

[8] Fan Dengping, Lin Zheng, Zhang Jiaxing, Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020: 463-473.

[9] Ran Ju, Ge Ling, Ge Wenjing, T. Ren, *et al.* Depth saliency based on anisotropic center-surround difference [C]// IEEE International Conference on Image Processing, 2014: 1115-1119.

[10] Hou Wenpeng, Li Bing, Wei Huaxiong, *et al.* RGBD salient object detection: A benchmark and algorithms [C]// Proceedings of the European Conference on Computer Vision (ECCV) , 2014: 92-109.

[11] Li Nianyi, Ye Jingwei, Yu Ji, et al Saliency detection on light field [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2806-2813.

[12] Piao Yongrui, Ji Wei, Li Jingjing, Zhang et al Depth-induced multiscale ecurrent attention network for saliency detection, [C]// The IEEE International Conference on Computer Vision, 2019: 7254-7263.

[13] Chen Hao, Li Youfu, Progressively complementarity-aware fusion network for RGB-D salient object detection [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3051-3060.

[14] Jiang Bing, Zhou Zhi, Wang Xing, et al cmSalGAN: RGB-D Salient Object Detection With Cross-View Generative Adversarial Networks [J]. IEEE Transactions on Multimedia, 2021: 1343-1353.

[15] Zhai Yinjie, Fan Dengping, Yang Jufeng, Bifurcated backbone strategy for RGB-D salient object detection [J]. IEEE Transactions on Image Processing, 2021: 8727-8742.

[16] Ji Wei, Li Jingjing, and Zhang Miao, Accurate RGB-D Salient Object Detection via Collaborative Learning [C]// Proceedings of the European Conference on Computer Vision (ECCV) , 2020: 52-69.

[17] Zhao Xiaoqi, Zhang Lihe, Pang Youwei, et al A Single Stream Network for Robust and Real-Time RGB-D Salient Object Detection [C]// Proceedings of the European Conference on Computer Vision (ECCV) , 2020: 646-662.

[18] Zhang Zhao, Lin Zheng, Xu Jun, Bilateral attention network for rgb-d salient object detection [J]. IEEE Transactions on Image Processing, 2021: 1949-1961.

[19] Li Gongyang, Liu Zhi, Ye Linwei, et al Cross-Modal Weighting Network for RGB-D Salient Object Detection [C]// Proceedings of the European Conference on Computer Vision (ECCV) , 2020: 665-681.

[20] Fan Dengping, Lin Zheng, Zhang Jiaxiang, et al, Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks, [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020: 2075-2089.

[21] Zhao Jiaxing, Cao Yang, Fan Dengping, *et al.* Contrast prior and fluid pyramid integration for RGBD salient object detection, [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) , 2019: 3927-3936.

[22] Li, Guanbing, and Yu Yizhou, Visual saliency based on multiscale deep features, [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2015: 5455-5463.

[23] Wu Zhe, Su Li, and Huang Qingming, Cascaded partial decoder for fast and accurate salient object detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2019: 3907-3916.

[24] Liu Zhenyi, Shi Song, Zhao Peng, et al Salient object detection for RGB-D image by single stream recurrent convolution neural network [C]// Neurocomputing, 2019: 46-57.

[25] Wu Junwei, Zhou Wujie, Luo Ting, *et al.* Multiscale multilevel context and multimodal fusion for RGB-D salient object detection, [C]// Signal Processing, 2021.

[26] 王豪聪，张松龙，彭力．融合边界信息和颜色特征的显著性区域检测 [J]. 计算机工程与应用, 2019, 55 (3): 179-183. (Wang Haocong, ZhangSonglong, Peng Li. Salient region detection based on fusion of boundary information and color features [J]. Computer engineering and Application, 2019, 55 (3): 179-183.)

[27] 翟继友，屠立忠，庄严．边界先验和自适应区域合并的显著性检测 [J]. 计算机工程与应用, 2018, 54 (6): 178-182. (Zhai Jiyou, Tu Lizhong, Zhuang yan. Significance detection of boundary a priori and adaptive region merging [J] Computer engineering and Application, 2018, 54 (6): 178-182.)